

## 2 Bodové a intervalové rozdělení četností

### 2.1 Jednorozměrné bodové rozdělení četností

#### Dataset 1: Porodní hmotnost novorozenců

Máme k dispozici údaje o porodní hmotnosti novorozenců z okresní nemocnice získané v období jednoho roku a současně máme k dispozici údaje o počtu starších biologických sourozenců novorozence, pohlaví novorozence a vzdělání matky (Alánová, 2008; soubor 17-anova-newb.txt).

#### Popis proměnných v datasetu 1:

- edu.M – vzdělání matky (1 – základní, 2 – střední bez maturity, 3 – střední s maturitou, 4 – vysokoškolské);
- prch.N – počet biologických starších sourozenců (0–2);
- sex.C – pohlaví dítěte (m – muž, f – žena);
- weight.C – porodní hmotnost dítěte (g);
- weight.K – porodní hmotnost dítěte (1 = nízká (nižší než 2 500 g), 2 = norma (2 500 – 4 200 g), 3 = vysoká (větší než 4 500 g))

#### Příklad 2.1. Načtení datového souboru

Načtěte dataset 17-anova-newb.txt do proměnné `data` a vypište prvních 5 řádků z načteného souboru. Zjistěte, zda soubor obsahuje neznámé (NA) hodnoty a pokud ano, tak je odstraňte. Potom zjistěte dimenzi datové tabulky `data`.

#### Řešení příkladu 2.1

	edu.M	prch.N	sex.C	weight.C	weight.K
1	2	0	m	3470	2
2	2	0	m	3240	2
3	2	0	f	2980	2
4	1	0	m	3280	2
5	3	0	m	3030	2

1  
2  
3  
4  
5  
6

Načtená datová tabulka obsahuje údaje o znacích: vzdělání matky (`edu.M`), počet starších sourozenců novorozence (`prch.N`), pohlaví novorozence (`sex.C`), porodní hmotnost novorozence (`weight.C`) a kategoriální porodní hmotnost novorozence (`weight.K`). Datový soubor obsahuje celkem ..... NA hodnot. Tabulka `data` má po odstranění NA hodnot celkem ..... řádků a ..... sloupců. V tabulce jsou tedy po odstranění NA hodnot uloženy údaje o ..... **objektech**, přičemž u každého objektu máme záznamy o ..... **znacích**.

#### Příklad 2.2. Úprava datového souboru

Upravte označení jednotlivých variant kategoriálního znaku *porodní hmotnost* tak, aby bylo na první pohled zřejmé, jakou hmotnost novorozenec má (1 = nízká, 2 = norma, 3 = vysoká). Analogicky upravte označení jednotlivých variant znaku *vzdělání matky* (1 – ZS, 2 – SS, 3 – SSm, 4 – VS).

#### Řešení příkladu 2.2

	edu.M	prch.N	sex.C	weight.C	weight.K
1	SS	0	m	3470	norma
2	SS	0	m	3240	norma
3	SS	0	f	2980	norma
4	ZS	0	m	3280	norma
5	SSm	0	m	3030	norma
6	SS	1	m	3650	norma

7  
8  
9  
10  
11  
12  
13

### Příklad 2.3. Variační řada

Vytvořte variační řadu znaku  $X = \text{vzdělání matky}$  a variační řadu kategoričkého znaku  $Y = \text{porodní hmotnost novorozence}$ .

#### Řešení příkladu 2.3

Zaměříme se nejprve na znak  $X = \text{vzdělání matky}$ . Znak má celkem čtyři varianty: ....., ....., ..... a ..... . Variační řada je tabulka obsahující pro každou ( $j$ -tou) variantu znaku  $X$  (a) absolutní četnost ..... ; (b) relativní četnost .....; (c) absolutní kumulativní četnost .....; (d) relativní kumulativní četnost .....

	n <sub>j</sub>	p <sub>j</sub>	N <sub>j</sub>	F <sub>j</sub>
ZS	347	0.2719	347	0.2719
SS	424	0.3323	771	0.6042
SSm	425	0.3331	1196	0.9373
VS	80	0.0627	1276	1.0000

14  
15  
16  
17  
18

**Interpretace výsledků:** Datový soubor obsahuje údaje o celkovém počtu ..... novorozenců, přičemž v 347 případech (27.19%) bylo nejvyšší dosažené vzdělání matky ....., v ..... případech (..... %) bylo nejvyšší dosažené vzdělání matky středoškolské bez maturity, apod. Celkem ..... (..... %) matek novorozenců v datovém souboru získalo středoškolské vzdělání bez maturity nebo nižší, celkem 1196 (93.73%) matek novorozenců získalo ..... nebo ..... vzdělání.

Zaměříme se nyní na znak  $Y = \text{porodní hmotnost novorozence}$ . Protože variační řadu má smysl sestřojovat pouze pro kategoriální / spojité znaky, použijeme k vytvoření variační řady proměnnou weight.C / weight.K. Znak  $Y$  má ..... varianty: nízká porodní hmotnost, norma a vysoká porodní hmotnost.

	n <sub>j</sub>	p <sub>j</sub>	N <sub>j</sub>	F <sub>j</sub>
nizka	240	0.1881	240	0.1881
norma	993	0.7782	1233	0.9663
vysoka	43	0.0337	1276	1.0000

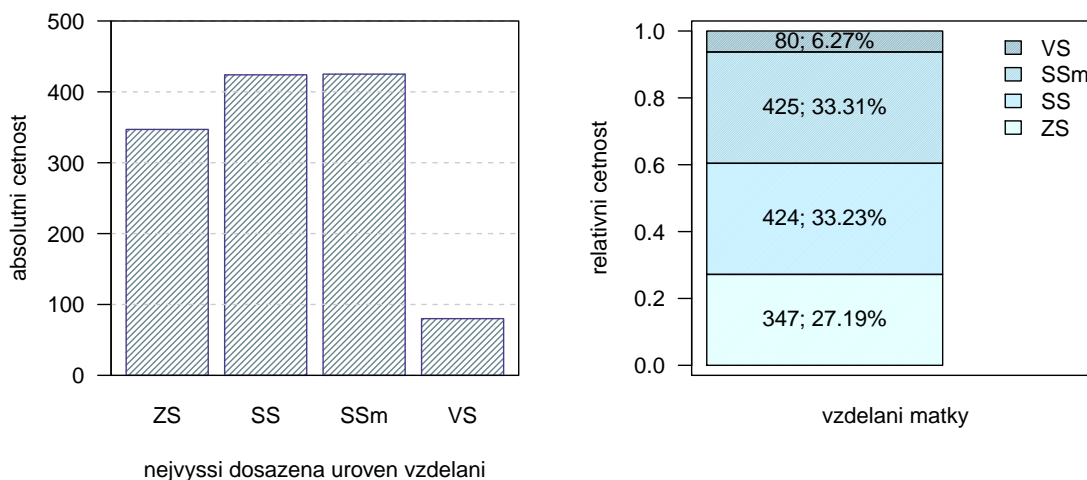
19  
20  
21  
22

**Interpretace výsledků:** Porodní hmotnost novorozenců v datovém souboru se v ..... případech (..... %) pohybovala v normě. Celkem ..... novorozenců (..... %) mělo porodní hmotnost nižší nebo rovnou normě a ..... novorozenců (..... %) mělo porodní hmotnost vysokou, v normě, nebo nižší.

### Příklad 2.4. Sloupcový graf absolutních a relativních četností

Nakreslete sloupcový graf absolutních četností a sloupcový graf relativních četností pro znak  $X = \text{vzdělání matky}$ .

#### Řešení příkladu 2.4



## Dvourozměrné bodové rozdělení četností

### Příklad 2.5. Kontingenční tabulka absolutních a relativních simultánních četností

Zaměřme se nyní na oba znaky  $X = \text{vzdělání matky}$  a  $Y = \text{porodní hmotnost novorozence}$  najednou. Z předchozího textu víme, že znak  $X$  má čtyři varianty, znak  $Y$  má tři varianty. Celkem tedy můžeme získat  $4 \cdot 3 = 12$  různých kombinací variant znaků  $X$  a  $Y$ . Sestrojte kontingenční tabulku simultánních absolutních četností a kontingenční tabulku simultánních relativních četností znaků  $X$  a  $Y$ .

### Řešení příkladu 2.5

Kontingenční tabulka simultánních absolutních četností bude tabulka o velikosti  $(4 + 1) \times (3 + 1) = 5 \times 4$  ve tvaru

	nizka	norma	vysoka	suma
ZS	$n_{11}$	$n_{12}$	$n_{13}$	$n_{1.}$
SS	$n_{21}$	$n_{22}$	$n_{23}$	$n_{2.}$
SSm	$n_{31}$	$n_{32}$	$n_{33}$	$n_{3.}$
VS	$n_{41}$	$n_{42}$	$n_{43}$	$n_{4.}$
suma	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n$

kde  $n_{jk}$ ,  $j = 1, \dots, 4$  a  $k = 1, \dots, 3$  je *simultánní absolutní četnost*  $j$ -té varianty znaku  $X$  a  $k$ -té varianty znaku  $Y$ ,  $n_{j.}$  (resp.  $n_{.k}$ ) je *marginální absolutní četnost*  $j$ -té varianty znaku  $X$  (resp.  $k$ -té varianty znaku  $Y$ ) a  $n$  je celkový počet objektů v datovém souboru.

#### Kontingenční tabulka simultánních absolutních četností

	nizka	norma	vysoka	suma
ZS	75	264	8	347
SS	79	325	20	424
SSm	73	341	11	425
VS	13	63	4	80
suma	240	993	43	1276

23  
24  
25  
26  
27  
28

**Interpretace výsledků:** V datovém souboru se vyskytuje celkem 75 novorozenců, kteří mají ..... porodní hmotnost a jejichž matka má ..... vzdělání, a ..... novorozenců, jejichž porodní hmotnost je v normě a jejichž matka má středoškolské vzdělání s maturitou. Celkem 80 novorozenců se narodilo matkám s ..... vzděláním.

#### Kontingenční tabulka simultánních relativních četností

	nizka	norma	vysoka	suma
ZS	0.0588	0.2069	0.0063	0.2719
SS	0.0619	0.2547	0.0157	0.3323
SSm	0.0572	0.2672	0.0086	0.3331
VS	0.0102	0.0494	0.0031	0.0627
suma	0.1881	0.7782	0.0337	1.0000

29  
30  
31  
32  
33  
34

**Interpretace výsledků:** V datovém souboru se vyskytuje celkem 5.88 % novorozenců, kteří mají ..... porodní hmotnost a jejichž matka má ..... vzdělání. V datovém souboru se vyskytuje celkem .....% novorozenců, jejichž porodní hmotnost je v normě a jejichž matka má středoškolské vzdělání s maturitou. Celkem 3.37 % novorozenců v datovém souboru má ..... porodní hmotnost.

### Příklad 2.6. Kontingenční tabulka řádkově a sloupcově podmíněných relativních četností

Zaměřte se nyní opět na oba znaky  $X = \text{vzdělání matky}$  a  $Y = \text{porodní hmotnost novorozence}$  najednou. Vytvořte kontingenční tabulku řádkově podmíněných relativních četností a kontingenční tabulku sloupcově podmíněných relativních četností.

#### Řešení příkladu 2.6

*Kontingenční tabulka řádkově podmíněných relativních četností*

	wei		
edu	nizka	norma	vysoka
ZS	0.2161	0.7608	0.0231
SS	0.1863	0.7665	0.0472
SSm	0.1718	0.8024	0.0259
VS	0.1625	0.7875	0.0500

35  
36  
37  
38  
39  
40

**Interpretace výsledků:** Ze všech novorozenců v datovém souboru, jejichž matka má dokončené středoškolské vzdělání zakončené maturitou, má 17.18% ..... porodní hmotnost a 2.59% ..... porodní hmotnost. Ze všech novorozenců v datovém souboru, jejichž matka má dokončené vysokoškolské vzdělání, má .....% **nízkou** porodní hmotnost a .....% **vysokou** porodní hmotnost.

*Kontingenční tabulka sloupcově podmíněných relativních četností*

	wei		
edu	nizka	norma	vysoka
ZS	0.3125	0.2659	0.1860
SS	0.3292	0.3273	0.4651
SSm	0.3042	0.3434	0.2558
VS	0.0542	0.0634	0.0930

41  
42  
43  
44  
45  
46

**Interpretace výsledků:** Ze všech novorozenců v datovém souboru, jejichž porodní hmotnost byla nízká, se 31.25% narodilo matkám s ukončeným ..... vzděláním. Ze všech novorozenců v datovém souboru, jejichž porodní hmotnost byla v normě, se .....% se narodilo matkám s dokončeným středoškolským vzděláním bez maturity.

## 2.2 Jednorozměrné intervalové rozdělení četností

### Dataset 2: Délkově-šířkové rozměry lebky egyptské populace

Z archivních materiálů (Schmidt, 1888; soubor 01-one-sample-mean-skull-mf.txt) máme k dispozici původní kranio-metrické údaje o délce a šířce mozkovny a ze starověké egyptské populace.

#### Popis proměnných v datasetu 2:

- id – pořadové číslo;
- pop – populace (egant – egyptská starověká);
- sex – pohlaví (m – muž, f – žena);
- skull.L – největší délka mozkovny (mm), t.j. přímá vzdálenost kranio-metrických bodů *glabella* a *opisthocranium*;
- skull.B – největší šířka mozkovny (mm), t.j. vzdálenost obou kranio-metrických bodů *euryon*.

#### Příklad 2.7. Načtení datového souboru

Načtete dataset 01-one-sample-mean-skull-mf.txt a vypište první čtyři řádky z načteného souboru. Prozkoumejte, zda soubor obsahuje neznámé hodnoty a případně je ze souboru odstraňte. Potom zjistěte dimenzi datové tabulky.

#### Řešení příkladu 2.7

	id	pop	sex	skull.L	skull.B
1	416	egant	m	188	145
2	417	egant	m	172	139
3	420	egant	m	176	138
4	421	egant	m	184	128

47  
48  
49  
50  
51

V datovém souboru se vyskytuje celkem ..... neznámých (NA) hodnot.

Po odstranění na pozorování nám zůstala datová tabulka o velikosti ..... řádků a ..... sloupců. Celkem tedy máme údaje o 325 ..... přičemž pro každý objekt máme ..... identifikační proměnnou id a údaje o ..... znacích: populaci (pop), pohlaví skeletu (sex), největší délce mozkovny (skull.L) a největší šířce mozkovny (skull.B).

### Příklad 2.8. Histogram a krabicový diagram

V následující analýze se zaměříme primárně na znak  $X =$  největší šířka mozkovny u skeletů mužského pohlaví. Proveďte prvotní náhled na znak  $X =$  největší šířka mozkovky u mužů pomocí (a) histogramu; (b) krabicového diagramu.

#### Řešení příkladu 2.8

Celkem máme údaje o největší šířce mozkovny u ..... mužských skeletů. Hodnoty největší šířky mozkovny v datovém souboru se pohybují v rozmezí .....-..... mm.

Jelikož je sledovaný znak  $X$  spojitého typu, je potřeba naměřené hodnoty roztrždit do stejně dlouhých tzv. *třídících intervalů*. V praxi to znamená, že vytvoříme intervaly pokrývající svým rozsahem celou reálnou osu, tj.

$$(\infty; u_1), (u_1; u_2), \dots, (u_r; u_{r+1}), (u_{r+1}; \infty),$$

kde  $(u_j; u_{j+1})$ ,  $j = 1, \dots, J$  je  $j$ -tý třídící interval. Krajní intervaly  $(\infty; u_1)$  a  $(u_{r+1}; \infty)$  jako třídící intervaly neuvažujeme, nikdy neobsahují žádné pozorování a slouží jako doplnění celé reálné osy. Počet třídících intervalů se mění v závislosti na počtu pozorování, které máme k dispozici. Přesný počet třídících intervalů  $r$  v konkrétním případě stanovíme pomocí tzv. Sturgesova pravidla

$$r \approx 1 + 3.3 \log_{10} n. \quad (1)$$

Podle Sturgesova pravidla je optimální počet třídících intervalů pro znak  $X =$  největší šířka mozkovny roven ..... Minimální naměřená hodnota znaku  $X$  je ....., maximální hodnota je ..... Rozsah hodnot mezi minimální a maximální hodnotou je .....

Optimální šířka třídícího intervalu pro znak  $X$  je ..... mm. Vynásobíme-li počet třídících intervalů optimálním rozsahem jednoho intervalu, zjistíme, že rozsah třídících intervalů je  $9 \times 3 = 27$ . Rozsah hodnot 124–149 je však pouze 25. Proto dolní hranici prvního třídícího intervalu  $u_1$  stanovíme jako 123,  $u_2 = 126, \dots, u_9 = 150$ .

